Madeira
Swiss Journal of Economics and Statistics (2022) 158:15
https://doi.org/10.1186/s41937-022-00093-5

Swiss Journal of Economics and Statistics

ORIGINAL ARTICLE Open Access

Partial identification of nonlinear peer effects models with missing data

Carlos Madeira*

Abstract

This paper examines inference on social interactions models in the presence of missing data on outcomes. In these models, missing data on outcomes imply an incomplete data problem on both the endogenous variable and the regressors. However, getting a sharp estimate of the partially identified coefficients is computationally difficult. Using a monotonicity property of the peer effects and a mean independence condition of individual decisions on the missing data, I show partial identification results for the binary choice peer effect model. A Monte Carlo exercise then summarizes the computational time and the accuracy performance of the interval estimators under some calibrations.

Keywords: Social interactions, Binary choice, Partial identification

JEL Classification: C25, C31, Z13

1 Introduction

In models of social interactions, the individual behavior depends both on individual characteristics and on aggregate characteristics of members of the group of which the agent is a member (Advani & Malde, 2018), integrating sociological concepts and economic thinking (Blume et al. 2010). Important applications of peer effects models have been developed for education (Sacerdote, 2001, Cipollone & Rosolia, 2007, Lalive & Cattaneo, 2009, Sojourner, 2013, Ammermueller & Pischke, 2009, Madeira, 2018), health behaviors (Bruhin et al., 2020, Bailey et al., 2021), employment (Roth, 2020) or migration (Slotwinski et al., 2019).

This work analyzes inference on nonlinear peer effects models in the presence of missing data. There are many situations (for example, drug use, teenage risk profiles, sexual behavior) where respondents might not be willing to reveal their personal experience, creating problems of missing data in the study of social interactions in these settings. Most social interaction studies use the average outcome of each group as an explanatory variable; therefore, missing outcome data imply that we face both a problem of missing outcome values and an undetermined regressor, aggravating the identification problem. It is, therefore, important to extend the robustness of the social interaction estimators to scenarios of missing data.

In the linear case, Manski (1993, 2000) showed that it is difficult to distinguish between the effects of endogenous social interactions and the impact of measures of exogenous group quality. Several works analyze identification of peer effects in the linear case (Advani & Malde, 2018, Sojourner, 2013, Ammermueller & Pischke, 2009). These works analyze partial and point identification of the linear peer effects model with missing data on outcomes. Sojourner (2013) shows that if individuals are randomly assigned to each group; then, it is possible to point-identify the true coefficient for the peer effects variable. Ammermueller and Pischke (2009) show that missing data on peers create measurement error for the group variables and using an analysis similar to Hausman (2001) find upper and lower bounds for the true peer effect coefficient of the linear model. The authors then

I would like to express my enormous debt to Elie Tamer, Chuck Manski, Orazio Attanasio, plus seminar participants at Northwestern University and the Econometric Society World Congress. Financial support from Fundação Calouste Gulbenkian is gratefully acknowledged. Comments are welcome at cmadeira@bcentral.cl. All errors are my own.

*Correspondence: cmadeira@bcentral.cl

Central Bank of Chile, Agustinas, 1180 Santiago, Chile

apply an instrument for the peer effects variable to obtain point identification.

However, several economic decisions such as discrete choices require nonlinear models (Blume et al., 2010). Nonlinear settings for peer effects include smoking behavior (Krauth, 2006), high school truancy, cell phone ownership (Kooreman & Soetevent, 2007) or college life (Sacerdote, 2001). Brock and Durlauf (2007) present a very general model of peer effects in a discrete choice setting, showing that it is possible to identify asymptotically both exogenous and endogenous peer effects under the assumption of random group assignment and no missing data.

I extend the identification results of Brock and Durlauf (2002, 2007) to the case of missing outcomes. Using an incomplete data approach proposed by Horowitz and Manski (2006), it is possible to get sharp bounds for the coefficients of this model with missing data, but this method can be time-consuming for larger peer groups. Therefore, I propose an estimator to obtain non-sharp bounds for this model based on Manski and Tamer (2002) interval regressors' approach. My suggested approach extends the interval regressors approach of Manski and Tamer (2002) by showing that it can easily be extended for a case with both interval regressors and missing outcomes. If a discrete choice model verifies three important properties—interval values (I), mean independence (MI), monotonicity (M)—then it is possible to obtain non-sharp bounds for the true coefficients of the model. The interval values (I) regressor assumption is trivially satisfied by discrete choice models with peer effects, since the average of the discrete choices in a peer group is bounded between 0 and 1. The mean independence (MI) is also quite natural in the peer effects model, since it implies that the width of the identification interval does not matter if one conditions on the true value of the average outcome. This assumption appears natural if the agents know the true values of the average choices in their peer groups even if the econometrician only observes the group with some missing data. The third assumption, monotonicity (M), implies that the average outcome of each agent is increasing with the average group outcome. This assumption is trivially satisfied in the parametric discrete choice models, and it can also be consistent with many semi-parametric or nonparametric models. A minimum distance estimator is proposed. I also propose a bootstrap method to estimate confidence intervals for the true coefficients. A similar estimator can be easily applied to any parametric model with missing outcomes and interval regressors.

I then show a set of Monte Carlo exercises with fully observed information to characterize the accuracy of the peer effects estimators even if the identification assumptions are satisfied. The Monte Carlo exercises include a wide range of different group sizes and different sample sizes for both the logit and the linear case. The Monte Carlo simulations include estimators for the cases of closed peer groups (groups in which all members are peers of each other) and non-closed groups (with each individual having peers from outside the group). Furthermore, I consider the case in which the individual is part of his own peer group and the case in which the individual is not part of its own peer group. The linear case is only shown for non-closed groups (which is required for identification, as shown in Bramoullé, Djebbari and Fortin 2009).

I then apply the Manski–Tamer and Horowitz–Manski estimators to the logit peer effects models in the presence of missing outcomes. The results show that the Manski–Tamer estimator can be hundreds of times faster than the Horowitz–Manski estimator even with just a few missing values such as 10 missing outcomes. The computation time of the Horowitz–Manski estimator could be much larger with a few additional missing observations.

This work focuses on the case in which missing information on missing outcomes also implies missing information or an interval regressor for the peer effect in order to be clear about this effect. This approach could also be easily generalized to other cases that also include missing control variables for the peer group members and which would also imply missing regressors or interval regressors. The case for other missing regressors would merely imply more combinations of possible datasets for the missing values for the Horowitz and Manski (2006) and additional interval regressors for the Manski and Tamer (2002) approaches suggested in this article.

This article is organized as follows: Section 2 shows how the interval regressors approach of Manski and Tamer (2002) can be easily extended for a case that also has missing outcomes. Section 3 explains the calibration of the Monte Carlo exercises. Section 4 then summarizes the Monte Carlo results in the absence of missing data. The section starts by showing that the exogenous coefficients (given by the constant, exogenous variable affecting individual behavior, contextual effects group variable) have a fast convergence to the true parameter values, whether the model has endogenous peer effects or not. The same simulations show that the endogenous peer effect coefficient has a much slower convergence to its true value, presenting a high bias and standard deviation, even without any missing data. Section 5 shows the Monte Carlo exercises with missing data, analyzing the performance of the Horowitz–Manski and Manski–Tamer approaches. The results show that the Horowitz and Manski (2006) approach presents a considerable computational time. The section also summarizes the

estimated interval results for all the coefficients, including both the endogenous peer effects parameter and the exogenous coefficients parameters. Finally, Sect. 6 summarizes the main results and an appendix shows the proofs of the main propositions.

2 Identification of discrete choice models with peer effects

2.1 A parametric discrete choice model

Let $y_i \in \{0,1\}$ represent individual i outcomes, $X_i \in R^K$ the individual exogenous variables, $g = 1, ..., G$ denotes the groups, and $Y_g \in R^Q$ is the set of exogenous variables for each group. I represent average group behavior as $p_{i,g} = \frac{1}{n_{g(i)}} \sum_{j=1, j\in g(i)}^{n} 1(y_j = 1)$, where $n_{g(i)} = \sum_{j=1}^{n} 1(j \in g(i))$ is the number of people in agent i's group.

Brock and Durlauf (2002) presented a parametric multinomial model of choice in the presence of social interactions, giving conditions for identification in the presence of fully observed data. Individual choice is determined by latent utility, $V_i = h_i + Jp_{i,g(i)} - \epsilon_i$. The term ϵ_i represents an idiosyncratic term (such as an individual taste factor) unobserved to the econometrician. ϵ_i has a known monotonic parametric distribution, $F_\epsilon(.)$. In this specification, h_i represents the components of utility affected only by exogenous variables, $h_i = k + bX_i + dY_{g(i)}$. The observable group variables $Y_{g(i)}$ for the contextual peer effects, correlated group effects or neighborhood variables (Manski, 1993) can include the mean values of the individual variables of the other group members. One can further specify $Y_{g(i)} = \frac{1}{n_{g(i)}} \sum_{j=1, j\in g(i)}^{n} X_j$ in the case of individuals that are part of their own peer group, or in alternative, $Y_{g(i)} = \frac{1}{n_{g(i)} - 1} \sum_{j=1, j\in g(i), j\neq i}^{n} X_j$ in the case in which the individual i is excluded from its own peer effect. In this model, the probability of choosing a positive outcomes is given by:

$$\begin{aligned} \Pr(y_i = 1) &= \Pr(V_i \geq 0) = \Pr(\epsilon_i \leq h_i + Jp_{i,g}) \\ &= F_\epsilon(h_i + Jp_{i,g}). \end{aligned} \tag{1}$$

It is possible that several values of $p_{i,g}$ might solve expression (1) due to multiple equilibria corresponding to self-consistent behaviors in the population (Brock & Durlauf, 2002).

This discrete choice model is quite parsimonious and includes all the main features of peer effects models. The term Y_g is usually interpreted as "contextual group effects" (Manski, 1993), meaning the gains each member of the group has due to exogenous characteristics of the group. For example, students of a certain school could be doing well because the school has good facilities and teachers. The term $p_{i,g}$ represents the "endogenous group effect" since it represents the feedback effect that group performance has on each one of its members. In this case, students of a certain school may be more likely to apply for college because the other students are also applying.

Brock and Durlauf (2002) show that this model is point-identified by assuming two conditions:

(i) $X_i, Y_{g(i)}, p_{i,g}$ are not collinear and Y_g has unbounded support;

(ii) ϵ_i are independent and identical distributed across individuals and are independent of X_i and $Y_{g(i)}$.

The independence assumption of ϵ_i can be relaxed when the group $g(i)$ of each individual is not entirely closed (for example, your neighbors have neighbors that are not neighbors of you). In this case, the "peers of your peers" provide extra variation that can be used for identification (Bramoullé et al. 2009). For now, this article will keep the assumption that the peer groups are closed and therefore $j \in g(i)$ implies that $i \in g(j)$.

2.2 Partial identification in the case of missing outcome information

If all the data are observed, one can estimate the parameters by using a maximum likelihood estimator (MLE):

$$\begin{aligned} \hat{\theta} \equiv (\hat{k}, \hat{b}, \hat{d}, \hat{J}) = \arg\max_{\theta} \sum_{i=1}^{N} y_i \ln(F_\epsilon(h_i + Jp_{i,g})) \\ + (1 - y_i) \ln(1 - F_\epsilon(h_i + Jp_{i,g})). \end{aligned} \tag{2}$$

Now assume $z \in \{0,1\}$ determines when y is unobserved or observed. For simplicity, I assume that X_i and $Y_{g(i)}$ are always observed, but the missing information on some outcomes y_i implies that $p_{i,g}$ is not point-identified, although $p_{i,g}$ can be bounded within a sharp interval. The number of missing observations in the sample is given by $n_{z=0} = \sum_{j=1}^{n} 1(z_j = 0)$. Let $m(i) = \sum_{j=1}^{i} 1(z_j = 0)$ denote the order of y_i in the sample of observations with missing values, with $m(i) = \varnothing$ if $z_i = 1$. I define $y_{z=0} \equiv \{y_m, (m = 1, ..., n_{z=0})\}$ as the vector collection of all missing outcome values in the sample. This vector belongs to the space given by $\Xi \equiv \{0,1\}^{n_{z=0}}$. Let $a \in \Xi$ be a feasible vector for the missing outcome values. I define $y_i^a = y_i$ if $z_i = 1$ and $y_i^a = a(m(i))$ if $z_i = 0$. In the same way, I define $p_{i,g}^a = \frac{1}{n_i} \sum_{j=1, j\in g(i)}^{n} 1(y_j^a = 1)$ as the group average outcome under the vector of missing values $y_{z=0} = a$.

Let $H(\theta)$ be the identified set of θ. Elements of this set can be identified by repeatedly plugging in feasible values for the missing data, $a \in \Xi$, and computing the parameters of interest. We can therefore study the set of values consistent with the observed data and the assumed model given any feasible distribution of the missing data. For a specific combination of the missing data values, $y_{z=0} = a \in \Xi$, one can estimate the coefficients $\hat{\theta}(a) = \arg\max_\theta \sum_{i=1}^{N} y_i^a \ln(F_\epsilon(h_i + Jp_{i,g}^a)) + (1 - y_i^a)\ln(1 - F_\epsilon(h_i + Jp_{i,g}^a))$ It is possible therefore to obtain $H(\theta)$ as $\hat{H}(\theta) \equiv \{\hat{\theta}(a),$ for all $a \in \Xi\}$. For finite samples, it is possible to obtain confidence intervals for the true coefficient parameters θ by using the bootstrap procedure of Imbens and Manski (2004). Another alternative is to get confidence intervals for the identified set, $H(\theta)$, by using a subsampling procedure described in Chernozhukov et al. (2007).

This plug-in strategy is fairly general and easy to implement. It can essentially work on any model that is proven to be identified and solvable. The difficulty of this approach, however, is that it is computationally demanding. Notice that if $\Pr(z = 0) > 0$, then the set of alternative values increases exponentially with N. Therefore, as N grows large this approach may require several alternative computations of $\hat{\theta}(a)$ to obtain a good estimate of $H(\theta)$.

Here, I show that less computationally intensive strategies are possible. Note that the Brock and Durlauf model is monotonic in the group outcome, $p_{i,g}$. Also, the regressor $p_{i,g}$ of each individual can be estimated to be inside an interval. Assume for simplicity we keep the group intervals $p_{i,g} \in [p_{i,g}^L, p_{i,g}^U]$ fixed, but we allow the values of each missing individual outcome $y_i \mid z_i = 0$ vary between $\{0, 1\}$. This approach gives us non-sharp bounds for the coefficients θ, since we are not taking into account that changing the individual $y_i \mid z_i = 0$ also has an effect on the interval of $p_{i,g}$. However, together with the monotonicity property of the peer effects model it is possible to specify a convenient estimator for these non-sharp bounds. For this reason, I generalize the approach of Manski and Tamer (2002) to the case of both missing outcomes and interval regressors.

2.3 Using monotonicity and mean independence assumptions

In the discrete choice case, outcomes y are bounded between 0 and 1. This guarantees the interval property (I) of the regressor, $p_{i,g}$, so we have $p_{i,g} \in [0, 1]$. It is easy to specify sharp bounds for the group average, $p_{i,g} = E[y \mid g(i)]$. Let us define $p_{g(i)}^L = E_L[y \mid g(i)] = E[y \mid g, z = 1]P(z = 1 \mid g)$ and $p_{g(i)}^U = E_U[y \mid g(i)] = E[y \mid g, z = 1]P(z = 1 \mid g) + P(z = 0 \mid g)$.

Then, the law of total probability gives us sharp bounds for the value of the group average, $p_{i,g}$, as expressed in Proposition (1):

Proposition 1 $p_{g(i)}^L \leq p_{i,g} \leq p_{g(i)}^U$.

Now, I denote $V_x^g = (p_{i,g}, Y_g, x)$, $W_x^g = (p_{g(i)}^L, p_{g(i)}^U, Y_g, x)$, $W_0^g = (p_{g(i)}^L, Y_g, x)$, *and* $W_1^g = (p_{g(i)}^U, Y_g, x)$. *I will show that under certain conditions it is possible to obtain a partial interval for* $E[y \mid V_x^g]$ *by using* $E[y \mid W_0^g]$ and $E[y \mid W_1^g]$.

Note that the model defined in expression (1) is mean independent of the missing data properties z_j. This guarantees the following mean independence (MI) property:

(MI) $F_\epsilon(. \mid W_x^g, p_{i,g}) = F_\epsilon(. \mid p_{i,g}, Y_g, x) = F_\epsilon(.)$, where the first equality is given by expression (1) and the second one is obtained after applying assumption (ii) which specifies ϵ_i as independent and identical distributed across individuals and independent of X_i and $Y_{g(i)}$.

The MI assumption is not testable, but seems realistic under many scenarios. If individuals actually observed average group behavior and act using this knowledge, then individual outcomes are not affected by missing data. For example, teenagers may know how many smokers or drug users exist in their group, even if the researcher does not. However, the MI assumption could be invalid if the individuals react more or less to the choices of their unreported peers.

Expression (1) also has the group endogenous effect, J, specified as a constant parameter. Since J is constant and $F_\epsilon(.)$ is monotonic, this guarantees the monotonicity (M) property of $E[y \mid V_x^g]$ in all its arguments. Therefore, the IMMI (interval, monotonicity, and mean independence property) described in Manski and Tamer (2002) holds for this model.

By applying the law of total probability, we get sharp bounds for $E[y \mid W_x^g]$.

Proposition 2 $E_L[y \mid W_x^g] \leq E[y \mid W_x^g] \leq E_U[y \mid W_x^g]$, *where* $E_L[y \mid W_x^g] = E[y \mid W_x^g, z = 1]\ P(z = 1 \mid W_x^g)$ *and* $E_U[y \mid W_x^g] = E[y \mid W_x^g, z = 1]P(z = 1 \mid W_x^g) + P(z = 0 \mid W_x^g)$.

Proposition 3 shows that it is possible to achieve exact identification of our parameters if and only if some groups in the population have no missing data at all.

Proposition 3 *Let* $\theta \equiv (k, b, d, J)$ *and* $c \equiv (c_1, c_2, c_3, c_4) \in C$. *Denote also* $h_i^c = c_1 + c_2' X_i + c_3' Y_{g(i)}$, $V_i^{c,U} = h_i^c + c_4 p_{g(i)}^U$ *and* $V_i^{c,L} = h_i^c + c_4 p_{g(i)}^L$. *Let*

$$V(c) = [(W_x^g) : F_\epsilon(h_i^c + c_4 p_{g(i)}^U) < E_L[y \mid W_x^g] \bigcup E_U[y \mid W_x^g] < F_\epsilon(h_i^c + c_4 p_{g(i)}^L)].$$

Then, θ is only identified relative to c if and only if $P[V(c)] > 0$.

Now, I characterize the identification region and present a minimum distance estimator for the parameters' identification set in the presence of missing outcome data. Lemma 1 forms the basis for the estimator. It shows that there is a parametric solution θ to the problem that fits within the bounds of the data moments ($E_L[y \mid W_x^g]$, $E_U[y \mid W_x^g]$), and therefore the solution is non-empty. It then characterizes the solution as being a convex interval and with a given expression that can be estimated. The importance of the interval solution being a convex region is important, because it implies that the problem is well behaved and many optimization methods for the econometric estimators only work with convex regions. Convexity implies that, for instance, if θ_L and θ_U (with $\theta_U \geq \theta_L$), then any other parameter given by $\theta^* = \theta_L(1-\alpha) + \alpha$, with $\alpha \in [0,1]$, is also a solution. This makes it convenient for optimization methods because it implies that there is a single convex identified interval, with$\theta_U \in [\theta_L, \theta_U]$. If the identified sets were not convex, then empirical researchers could find several unconnected regions with blanks in between. It would even make it difficult to determine if the entire identified set had been found by the researcher, since there could be more identified regions in other areas.

Lemma 1 *The identification region for θ is non-empty, convex and equivalent to*

$$H(\theta) = \arg\min_{c \in C} \int 1[F_\epsilon(W_1^g) < E_L[y \mid W_x^g]]\,[F_\epsilon(W_1^g) - E_L[y \mid W_x^g]]^2 + 1[F_\epsilon(W_0^g) > E_U[y \mid W_x^g]]\,[F_\epsilon(W_0^g) - E_U[y \mid W_x^g]]^2 dP(W_x^g),$$

where $F_\epsilon(W_{1,i}^g) = F_\epsilon(h_i + Jp_{g(i)}^U) = F_\epsilon(k + b'X_i + d'Y_{g(i)} + Jp_{g(i)}^U)$ *and* $F_\epsilon(W_{0,i}^g) = F_\epsilon(h_i + Jp_{g(i)}^L) = F_\epsilon(k + b'X_i + d'Y_{g(i)} + Jp_{g(i)}^L)$.

Proposition 4 gives the basic finding on identification of parametric regression models. It suggests that the identification region can be found by a minimum distance estimator $H_N(\theta)$, which uses the empirical analogs of the bounds $E_L^N[y \mid W_{x,i}^g]$ and $E_U^N[y \mid W_{x,i}^g]$ in order to find the identified set for θ. Note that it is relevant for the minimum distance optimizer to search for the parameter values that fit within the intervals given by $E_L^N[y \mid W_{x,i}^g]$ and $E_U^N[y \mid W_{x,i}^g]$. In general, the researcher cannot just obtain two estimators given by replacing the missing values with zeros or replacing the missing values with ones, because that could imply establishing that a correlation between $p_{g(i)}^U$ and the other variables (X_i and $Y_{g(i)}$) is either very low or very high in order to explain a large amount of zeros and ones. The minimum distance estimator must therefore search for the parameter values until the region that fits the empirical analogs is found.

Proposition 4 *A suggested estimator for the identification region $H(\theta)$ would be*

$$H_N(\theta) = \arg\min_{c \in C} \frac{1}{N} \sum_{i=1}^{N} 1[F_\epsilon(W_{1,i}^g) < E_L^N[y \mid W_{x,i}^g]]\,[F_\epsilon(W_{1,i}^g) - E_L^N[y \mid W_{x,i}^g]]^2 + 1[F_\epsilon(W_{0,i}^g) > E_U^N[y \mid W_{x,i}^g]]\,[F_\epsilon(W_{0,i}^g) - E_U^N[y \mid W_{x,i}^g]]^2$$

where $E_L^N[y \mid W_{x,i}^g]$ and $E_U^N[y \mid W_{x,i}^g]$ are consistent estimators of $E_L[y \mid W_{x,i}^g]$ and $E_U[y \mid W_{x,i}^g]$, respectively. It is possible to take into account finite-sample error in the estimates of these intervals by using the bootstrap technique described by Imbens and Manski (2004). *A similar identification strategy is possible for the semi-parametric peer effects model developed in Brock and Durlauf* (2007), *although such a discontinuous estimator does not have a convenient asymptotic distribution and therefore does not allow to obtain a finite-sample confidence interval by using the bootstrap method of Imbens and Manski* (2004).

2.4 Horowitz and Manski's estimator for functionals of incomplete data

Let us define a parametric estimator

$$\theta = \arg\min_{\theta} \sum_{i=1}^{N} f(y_i, \bar{y}_{i,g}, H_i)$$

with $H_i = (X_i, Y_{g(i)}, w_i)$ being the vector of control variables that are completely observed. Again assume $y_{i,z=1}$ is observed and $y_{i,z=0}$ is not observed. The estimator obtained from the true dataset can be expressed as:

$$\theta = \arg\min_{\theta} \sum_{i=1,z=1}^{N} f(y_{i,z=1}, \bar{y}_{i,g}$$
$$= \frac{1}{n_{g(i)}} \sum_{j=1}^{n_{g(i)}} y_{j,z=1} + y_{j,z=0}, H_i) + \sum_{i=1,z=0}^{N} f(y_{i,z=0}, \bar{y}_{i,g}$$
$$= \frac{1}{n_{g(i)}} \sum_{j=1}^{n_{g(i)}} y_{j,z=1} + y_{j,z=0}, H_i).$$

The econometrician does not observe the values of $y_{i,z=0}$, but it knows that each value of y belongs to a finite set Υ with V elements. One possible estimator can be obtained from one of the possible ways of imputing the missing dataset:

$$\theta_c = \arg\min_{\theta} \sum_{i=1,z=1}^{N} f(y_{i,z=1}, \bar{y}_{i,g}$$
$$= \frac{1}{n_{g(i)}} \sum_{j=1}^{n_{g(i)}} y_{j,z=1} + v_{j,z=0}, H_i) + \sum_{i=1,z=0}^{N} f(v_{i,z=0}, \bar{y}_{i,g}$$
$$= \frac{1}{n_{g(i)}} \sum_{j=1}^{n_{g(i)}} y_{j,z=1} + v_{j,z=0}, H_i),$$

with $v_{i,z=0} \in \Upsilon$ being a specified value for each possible missing observation i. Since there are V possibilities for each $y_{i,z=0}$, there are $MD = V^{N_{z=0}}$ possible datasets that could be validly imputed, with $N_{z=0} = \sum_{i=1}^{N} 1(z_i = 0)$ denoting the number of missing observations. Then, if one computes the estimator θ_c across all the possible imputations for $y_{i,z=0}$, the econometrician will find that $\theta \in \{\theta_1, ., \theta_c, .., \theta_{MD}\}$; therefore, a sharp interval for θ can be obtained as:

$$\theta \in [\inf\{\theta_1, ., \theta_c, .., \theta_{MD}\}, \sup\{\theta_1, ., \theta_c, .., \theta_{MD}\}]. \quad (3)$$

This sharp interval obtained from the infimum and supremum estimates obtained across all the possible realizations of the dataset can be easily applied to any parametric estimator (Horowitz & Manski 2006), such as the parametric discrete choice model or the linear model of peer effects exposed before. Note that the number of possible estimates MD increases rapidly with just a few missing observations. For instance, in the discrete choice setting ($V = 2$, since y can be 0 or 1), the number of possible datasets would reach $MD = 2^{15} = 32768$ possible datasets with just 15 missing observations.

In the appendix, I show that similar econometric approaches can be applied to linear social interactions' models that have bounded outcomes, assuming that identification is obtained through the observation of non-closed groups.

3 Monte Carlo exercises

3.1 The simulated models

The discrete choice model of social interactions for simulation $s = 1, .., S$ is obtained as follows. The discrete peer effects model of each simulation s is specified to be a logit, $\Pr(y_i(s) = 1) = \Lambda(k + bX_i(s) + dY_{g(i)}(s) + Jp_{i,g}(s))$, with $\Lambda(x) = \frac{\exp(x)}{1+\exp(x)}$. For simplicity, the simulations consider that the coefficients $\{k, b, d, J\}$ are constants for all simulations and that all the groups are the same, that is, $n_{g(i)} = n_g$. Each simulated observation is obtained with the specified set of coefficients: $k = -2$, $b = 1$, $d = 0.5$, $J = 0.5$. The observable control variables $X_i(s)$ and $Y_{g(i)}(s)$ are simulated as independent pseudo-standard normal numbers, with $X_i(s)$ having different values for each individual i in each simulation s and $Y_{g(i)}(s)$ having different values for each group g in each simulation s. For each simulation s, each observation is then obtained from pseudo-uniform numbers: $y_i(s) = 1(\varepsilon_i(s) \leq k + bX_i(s) + dY_{g(i)}(s) + Jp_{i,g}(s))$, with $\varepsilon_i(s)$ being pseudo-standard logistic random numbers with mean 0 and standard deviation $\pi/\sqrt{3}$.

The exercises consider an alternative with closed groups, with $p_{i,g}(s)$ given entirely by the endogenous decisions of the members of each group $g = 1, .., G$, and an alternative with non-closed groups with each individual i reporting a value $w_i(s)$ for the peer effect of the members outside the group. The exercise considers $w_i(s)$ as given by a pseudo-uniform number independent across i and s. Furthermore, the peer effects of the observed group members are considered in two versions with the first one considering the individual as parts of its own peer group, $\sum_{j=1,j\in g(i)}^{n} 1(y_j(s) = 1)/n_g$, while a second version considers that the individual is not part of its own peer effect $\sum_{j=1,j\in g(i),j\neq i}^{n} 1(y_j(s) = 1)/(n_g - 1)$. The reason for these two alternatives is that considering the individual as part of its own peer effect introduces an obvious problem, since $p_{i,g}(s)$ includes $y_i(s)$ which is a function of the unobserved idiosyncratic error $\varepsilon_i(s)$. Therefore, it is likely that estimators that consider individuals as part of their own peer effect should present a bias due to the control variable being correlated with the unobserved error (Wooldridge 2010).

Therefore, the variable $p_{i,g}(s)$ is implemented in four alternatives:

(i) Closed groups with individual i as part of his own peer group, $p_{i,g}(s) = \frac{\sum_{j=1,j\in g(i)}^{n} 1(y_j(s) = 1)}{n_g}$;

(ii) Closed groups with individuals excluded from their own peer group, $p_{i,g}(s) = \frac{\sum_{j=1, j\in g(i), j\neq i}^{n} 1(y_j(s)=1)}{n_g - 1}$;

(iii) Non-closed groups with individual i as part of his own peer group and with the outside peer group of the same size as the group g, $p_{i,g}(s) = \frac{n_g w_i(s) + \sum_{j=1, j\in g(i)}^{n} 1(y_j(s)=1)}{n_g + n_g}$;

(iv) Non-closed groups with individual i excluded from their own peer group and with the outside peer group of the same size as the group g, $p_{i,g}(s) = \frac{n_g w_i(s) + \sum_{j=1, j\in g(i), j\neq i}^{n} 1(y_j(s)=1)}{n_g + n_g - 1}$.

The Monte Carlo exercises consider several combinations of group size with $n_g = 5, 10, 25$ members and several numbers of groups with $G = 50, 100, 200, 500, 1000, 2500$. The total sample size in terms of individuals is given by $N = G \times n_g$. Since some exercises take a long time (in particular, the Horowitz–Manski estimator takes a longer time than the other with higher values of missing data), all the Monte Carlo exercises are done with just 50 simulations, $s = 1, .., S$, with $S = 50$.

To summarize the results from the Monte Carlo simulations, I denote θ as the vector with the true value of the parameters, $\theta = \{k, b, d, J\}$, while $\hat{\theta}_s$ denotes the estimate obtained in each simulation. The average estimate across all the simulations is obtained as $\bar{\theta} = \frac{\sum_s \hat{\theta}_s}{S}$. The mean bias is therefore computed as $\bar{\theta} - \theta$, while the standard deviation (STD) is given by $\sqrt{\frac{\sum_s (\hat{\theta}_s - \bar{\theta})^2}{S-1}}$ and the mean absolute deviation (MAD) is $\frac{\sum_s \left|\hat{\theta}_s - \theta\right|}{S}$. The mean absolute deviation (MAD) can be a better measure of the small sample performance of the estimators than the standard deviation (STD), especially because it is possible that some estimators have a considerable bias and the bias effect is not part of the standard deviation (STD).

All the Monte Carlo exercises were performed in a notebook with an Intel Core i7-9750H 2.60GHz, with 24.0 GB of RAM, 6 physical cores and 12 logical processors. The codes were implemented with a Stata 15.1 MP-6 software license. All the codes are publicly available in the Mendeley Data repository: https://data.mendeley.com/datasets/zsbxdmhtj9/1.

3.2 Calibrating the missing observations

The missing observations are specified in terms of the number of missing outcomes ($y_i(s)$) in each simulation s. I create independent pseudo-uniform numbers $zu_i(s)$, and then for each simulation s specify as missing observations those with the m lowest values of $zu_i(s)$. For simplicity, all the other variables are observed (for instance, a variable X for family education or house type could be observed from administrative data), except for the endogenous variable $y_i(s)$. I choose this option instead of a probability, because the Horowitz–Manski estimator would require a number of $MD = V^{N_{z=0}}$ possible datasets, with V being the possible values of $y_i(s)$ and $N_{z=0}$ being the number of missing observations. This implies that even a small number of observations such as 15 would reach $MD = 2^{15} = 32768$ possible datasets and a very large computational time. For this reason, I prefer to specify the number of missing values, rather than a probability of missing outcomes which would result in a random number of missing outcomes for each simulation. In the case of the logit model, I will show Monte Carlo exercises with 5 and 10 missing values.

4 Monte Carlo exercises without missing data

This section starts by presenting the results of the Monte Carlo exercises without missing data. Table 1 summarizes the mean bias, standard deviation (STD) and mean absolute deviation (MAD) of the estimated coefficients, excluding the J endogenous effect. I compare the logit model with only contextual group effects (that is, assuming $J = 0$) with the logit endogenous peer effects model with closed groups (as suggested in Brock and Durlauf (2002)), although with the individuals excluded from their own peer effect. The contextual effects only model can be seen as a more traditional model, since there is no endogenous control variable and no correlation among individuals apart from the observable group effect $Y_{g(i)}(s)$. The results show that the logit with only contextual effects converges quite quickly to the truth and the estimator presents accurate values even with just 5 members per group and 50 groups (therefore a total sample of 250 observations). However, the logit model with both endogenous and contextual effects also converges somewhat quickly toward the true values of the parameters. The same pattern appears with the logit model with non-closed groups, which is shown in Table 2.

Table 3 shows how important it is for estimation of the logit endogenous peer effects coefficient (J) excluding the individuals from their own peer group $g(i)$ and whether it is helpful or not to include peers from outside the group (non-closed groups, which are essential for identification of the linear model). Models 1 and 2 show the case of closed groups, with individuals excluded and included from their own peer group, respectively. Models 3 and 4 show the case of non-closed groups, with individuals excluded and included from their own peer group, respectively. The results show that it is very important to

Table 1 Bias, standard deviation and mean absolute deviation of the estimates of the logit discrete choice model with contextual group effects and the logit endogenous social interactions model with closed groups

Group size	No. of groups	Bias: $\bar{\theta}-\theta\left(\bar{\theta}=\frac{\sum_S \hat{\theta}_s}{S}\right)$			STD: $\sqrt{\frac{\sum_S(\hat{\theta}_s-\bar{\theta})^2}{S-1}}$			MAD: $\frac{\sum_S\left\|\hat{\theta}_s-\theta\right\|}{S}$			Average time (secs)
		k	*b*	*d*	*k*	*b*	*d*	*k*	*b*	*d*	
Logit with contextual group effects ($J=0$)											
5	50	−0.03	0.03	0.00	0.21	0.20	0.17	0.17	0.17	0.13	0.02
5	100	−0.03	0.01	0.04	0.18	0.18	0.14	0.13	0.11	0.11	0.02
5	200	−0.03	0.01	0.00	0.12	0.11	0.10	0.10	0.08	0.08	0.02
5	500	0.00	0.00	0.01	0.07	0.07	0.05	0.06	0.05	0.04	0.04
5	1000	0.00	0.01	0.00	0.05	0.04	0.05	0.04	0.03	0.04	0.07
5	2500	0.00	0.00	0.00	0.03	0.03	0.03	0.03	0.02	0.02	0.13
Logit endogenous social interactions model (closed groups)											
5	50	0.09	0.04	0.02	0.29	0.18	0.22	0.24	0.15	0.17	0.03
5	100	0.11	0.01	−0.03	0.24	0.17	0.12	0.20	0.14	0.10	0.03
5	200	0.02	0.03	−0.01	0.16	0.12	0.10	0.12	0.10	0.08	0.04
5	500	0.05	0.01	−0.01	0.10	0.06	0.06	0.09	0.05	0.05	0.07
5	1000	0.03	0.00	−0.01	0.07	0.06	0.03	0.06	0.04	0.03	0.08
5	2500	0.05	0.00	−0.03	0.04	0.03	0.02	0.05	0.03	0.03	0.14
10	50	0.05	0.00	−0.02	0.25	0.12	0.15	0.21	0.10	0.11	0.03
10	100	0.01	0.00	−0.02	0.19	0.09	0.10	0.15	0.07	0.08	0.03
10	200	0.00	−0.01	−0.01	0.12	0.07	0.08	0.10	0.06	0.07	0.04
10	500	0.00	−0.01	−0.03	0.09	0.05	0.04	0.07	0.04	0.05	0.07
10	1000	−0.01	0.00	−0.03	0.06	0.03	0.03	0.05	0.02	0.03	0.12
10	2500	−0.03	0.00	−0.02	0.04	0.02	0.02	0.04	0.02	0.02	0.25
25	50	0.02	0.01	0.04	0.26	0.08	0.12	0.21	0.06	0.09	0.03
25	100	−0.01	0.00	−0.03	0.18	0.07	0.08	0.14	0.05	0.06	0.05
25	200	−0.07	0.00	−0.04	0.09	0.05	0.05	0.09	0.04	0.06	0.07
25	500	−0.06	0.00	−0.03	0.07	0.04	0.04	0.08	0.03	0.04	0.14
25	1000	−0.05	0.00	−0.03	0.05	0.02	0.03	0.06	0.02	0.03	0.26
25	2500	−0.05	0.00	−0.03	0.03	0.01	0.01	0.06	0.01	0.03	0.59

exclude individuals from their own peer effect in order to estimate *J*, because models M2 and M4 present large values for the mean bias and mean absolute deviation (MAD), with such values falling slowly as the number of group members increases (the number of group member reduces the effect of the individual in its own group $\frac{1}{n_g}$, besides increasing the sample size) and with the number of groups (which increases the sample size). This shows it is not advisable in practice for empirical researchers to include individuals as part of their own peer group, even if the model is identified in theory. Both M1 and M3 present accurate estimations in the sense that both models exclude individuals from their own peer group. However, M3 also includes peer effects from outside the group $\left(p_{i,g}(s)=\frac{n_g w_i(s)+\sum_{j=1,j\in g(i),j\neq i}^{n} 1(y_j(s)=1)}{n_g+n_g-1}\right)$.

The Monte Carlo exercise reveals that including peer effects outside of the group (model M3) can increase the mean bias, standard deviation (STD) and mean absolute deviation (MAD) for small sample sizes, such as just 50 groups. However, the model with non-closed groups (M3) can present a lower bias for larger sample sizes, although with a larger standard deviation. It is only for large sample sizes (group size of 25 members and a number of groups of 500 or more, which implies a sample size equal or bigger than 12,500 observations) that the non-closed groups model M3 represents a lower mean absolute deviation relative to the closed group model M1. This makes sense, since the additional control variable (the outside peer effects $w_i(s)$) represents an additional source of identification, but it also increases the dispersion in individual and group outcomes.

Table 2 Bias, standard deviation and mean absolute deviation of the estimates of the logit endogenous social interactions model with non-closed groups. 50 Monte Carlo simulations. Individuals are excluded from their own peer group

Group size	No. of groups	Bias: $\bar{\theta} - \theta(\bar{\theta} = \frac{\sum_S \hat{\theta}_s}{S})$			STD: $\sqrt{\frac{\sum_S (\hat{\theta}_s - \bar{\theta})^2}{S-1}}$			MAD: $\frac{\sum_S \lvert\hat{\theta}_s - \theta\rvert}{S}$			Average time (secs)
		k	*b*	*d*	*k*	*b*	*d*	*k*	*b*	*d*	
Logit endogenous social interactions model (non-closed groups)											
5	50	0.05	0.02	0.02	0.42	0.19	0.23	0.34	0.15	0.18	0.03
5	100	0.01	0.01	−0.01	0.31	0.12	0.12	0.24	0.10	0.10	0.03
5	200	0.05	0.02	0.02	0.21	0.12	0.12	0.19	0.10	0.10	0.04
5	500	0.04	0.00	−0.01	0.11	0.06	0.06	0.09	0.05	0.04	0.07
5	1000	0.02	0.01	0.00	0.09	0.04	0.05	0.08	0.03	0.04	0.08
5	2500	0.02	0.00	0.00	0.06	0.03	0.02	0.05	0.02	0.02	0.14
10	50	0.06	0.03	0.04	0.43	0.16	0.15	0.34	0.12	0.12	0.03
10	100	0.00	0.04	0.02	0.24	0.11	0.11	0.20	0.09	0.09	0.03
10	200	0.02	−0.01	0.01	0.19	0.07	0.07	0.15	0.06	0.06	0.04
10	500	0.00	0.00	0.01	0.11	0.05	0.04	0.09	0.04	0.04	0.07
10	1000	−0.01	0.00	0.00	0.07	0.03	0.03	0.06	0.03	0.03	0.12
10	2500	−0.02	0.00	−0.01	0.05	0.02	0.02	0.04	0.02	0.02	0.25
25	50	0.01	0.02	0.01	0.26	0.08	0.10	0.22	0.07	0.08	0.03
25	100	0.01	0.01	0.00	0.18	0.07	0.06	0.13	0.06	0.05	0.05
25	200	−0.03	0.00	0.00	0.16	0.04	0.05	0.13	0.03	0.04	0.07
25	500	0.01	0.01	0.00	0.08	0.02	0.03	0.06	0.02	0.03	0.14
25	1000	−0.04	0.00	−0.01	0.05	0.02	0.02	0.05	0.02	0.02	0.25
25	2500	−0.02	0.00	0.00	0.03	0.01	0.01	0.03	0.01	0.01	0.62

5 Monte Carlo exercises with missing data

5.1 Time performance of the interval estimators

This section summarizes the Monte Carlo results of the Horowitz–Manski and Manski–Tamer type of estimators. Table 4 compares the average computational time of each estimator. Note that the 5 missing observations correspond to a very small probability of missing outcomes, ranging from just 0.01% in the large sample cases to a maximum of 2% for the lowest samples. In the case of 7 and 10 missing observations, the corresponding probability of missing outcomes ranges from 0.06% to 2.8% and 0.02% to 4%, respectively. These are very low probabilities of missing data, since it is quite common to find survey datasets with more than 4% of missing data. For the case of just 5 missing observations, the Horowitz–Manski for the logit model takes between 0.7 and 30.6 seconds for the average across all simulations, while the Manski–Tamer type of estimator takes between 0.6 and 19.6 seconds, which can be 50% faster in some cases. For 10 missing observations, the time performance difference among the two estimators grows much larger, with the Horowitz–Manski type of estimator taking between 23 and 966 seconds, while the Manski–Tamer estimator keeps about the same computational time as with just 5 missing observations, with an average time between 0.5 and 19.7 seconds. The conclusion is that the number of combinations required to compute the Horowitz–Manski type of estimator increases exponentially with the number of observations ($MD = V^{N_{z=0}}$), while for the Manski–Tamer the calculation remains similar even as the number of missing outcomes increases.

5.2 Intervals of the interval estimators

Now, I summarize the mean intervals across all simulations of the Horowitz–Manski and the Manski–Tamer around the true parameter values. Table 5 shows the mean intervals for the case of the logit model with 5 missing values. For the case of the parameters *k*, *b* and *d*, the Horowitz–Manski type of interval estimator almost always contains the true parameter value in its average interval, although the intervals can be large in small samples such as 50 groups. However, for the case of the endogenous peer effects coefficient *J*, the Horowitz–Manski type of estimator often gives a biased interval that does not contain the true parameter value, as shown for the simulations with group sizes of 10 and 25

Table 3 Bias, standard deviation and mean absolute deviation of the estimates for the endogenous effects coefficient (*J*) of the logit endogenous social interactions model

Group size	No. of groups	Bias: $\bar{\theta} - \theta(\bar{\theta} = \frac{\sum_s \hat{\theta}_s}{S})$				STD: $\sqrt{\frac{\sum_s (\hat{\theta}_s - \bar{\theta})^2}{S-1}}$				MAD: $\frac{\sum_s \lvert\hat{\theta}_s - \theta\rvert}{S}$			
		M1	M2	M3	M4	M1	M2	M3	M4	M1	M2	M3	M4
5	50	−0.32	6.68	−0.63	10.19	1.29	0.75	2.19	1.90	0.91	6.68	1.80	10.19
5	100	−0.10	6.72	0.04	9.52	0.84	0.60	1.55	1.39	0.60	6.72	1.22	9.52
5	200	0.01	6.70	−0.20	9.62	0.54	0.41	1.16	0.85	0.41	6.70	0.95	9.62
5	500	0.09	6.72	0.01	9.62	0.40	0.19	0.55	0.46	0.29	6.72	0.42	9.62
5	1000	0.10	6.67	−0.01	9.62	0.21	0.17	0.42	0.39	0.18	6.67	0.34	9.62
5	2500	0.11	6.67	0.07	9.66	0.14	0.12	0.28	0.27	0.15	6.67	0.23	9.66
10	50	−0.15	6.47	−0.51	7.67	1.23	0.88	2.35	1.53	0.95	6.47	1.90	7.67
10	100	0.03	6.52	−0.13	8.11	0.92	0.57	1.19	1.14	0.75	6.52	0.99	8.11
10	200	0.16	6.31	−0.01	8.05	0.62	0.40	1.07	0.78	0.54	6.31	0.83	8.05
10	500	0.19	6.26	0.03	8.09	0.36	0.25	0.62	0.45	0.32	6.26	0.48	8.09
10	1000	0.23	6.30	0.14	8.18	0.26	0.15	0.37	0.29	0.29	6.30	0.31	8.18
10	2500	0.30	6.30	0.22	8.15	0.14	0.09	0.25	0.23	0.31	6.30	0.27	8.15
25	50	−0.17	6.20	−0.02	5.27	1.28	0.80	1.54	1.27	1.05	6.20	1.28	5.27
25	100	0.15	6.12	−0.12	5.19	0.83	0.53	0.93	1.00	0.67	6.12	0.74	5.19
25	200	0.44	6.18	0.17	5.19	0.50	0.35	0.83	0.56	0.53	6.18	0.69	5.19
25	500	0.33	6.08	0.02	5.29	0.35	0.18	0.41	0.36	0.41	6.08	0.33	5.29
25	1000	0.35	6.09	0.27	5.23	0.27	0.13	0.26	0.35	0.38	6.09	0.31	5.23
25	2500	0.36	6.09	0.14	5.26	0.18	0.08	0.19	0.19	0.37	6.09	0.18	5.26

Model 1: Closed groups, with individuals excluded from their own peer group. Model 2: Closed groups, with individuals as part of their own peer group. Model 3: Non-closed groups, with individuals excluded from their own peer group. Model 4: Non-closed groups, with individuals as part of their own peer group. 50 Monte Carlo simulations

for samples with 500 groups or more. The Manski–Tamer always has a larger interval than the Horowitz–Manski, especially for small samples as 50 groups, but this difference becomes quite small for a number of groups of 100 or more. The bounds of the Horowitz–Manski and the Manski–Tamer estimators tend to be reasonably small for samples with 1000 or 2500 groups, although with a significant bias for the *J* parameter.

It is problematic that in a few cases the bounds of the Horowitz–Manski and Manski–Tamer estimators do not include the true parameter value for the endogenous peer effect parameter *J*. This happens only for large peer groups (a group size of 10 or 25 members) and only for a large number of groups (500 groups or more). It is not easy to clarify why this inconsistency of the interval estimators is happening, but the previous literature shows three factors that complicate the estimation of discrete choice models, particularly those with correlated observations. One factor is that all the nonlinear models (which includes the logit model) have a certain degree of bias in finite samples and this appears in the Monte Carlo exercises (Wooldridge, 2010). A second factor is that this small sample inconsistency of the discrete choice model is further exacerbated in settings with panel data (Heckman, 1981, Honoré & Tamer, 2006)[1]. A third factor is that the literature shows that misclassification of dependent variables in a discrete-response model causes inconsistent coefficient estimates (Hausman et al., 1998). This is a very close example to the setting of this paper, since the interval estimators work by trying several possible options for the missing outcomes and the endogenous group averages, which is in effect working with many samples that are misclassified and only a single sample that represents the true outcomes.

It also happens sometimes for the other parameters *k*, *b* and *d* that the lower bound $\bar{\theta}_{\min}$ excludes the true parameter value, but the estimated interval is always very close

[1] While panel data are not the same as peer effects, both cases are examples in which the observations are correlated among themselves through heterogeneity and endogeneity (Heckman, 1981, Honoré & Tamer, 2006). The heterogeneity comes from the random effect for the panel data and the contextual effect for the social interactions model. The endogeneity issue in these models comes from the dynamic effect of previous choices in the case of panel data and the effect of the endogenous peer choices in the social interactions model. Honoré and Tamer (2006) show that the dynamic discrete choice models are hard to identify; therefore, this should explain why the peer effects model is also harder to estimate as the group size grows larger, and therefore the observations become more correlated among themselves.

Table 4 Time performance of the Horowitz–Manski (HM) and Manski–Tamer (MT) interval estimators (with non-closed groups and individuals excluded from their own peer group)

Group size	No. of groups	5 missing observations			10 missing observations			7 missing observations		
		Prob. of missing	Average time (secs)		Prob. of missing	Average time (secs)		Prob. of missing	Average time (secs)	
		(in %)	Logit		(in %)	Logit		(in %)	OLS	
			HM	MT		HM	MT		HM	MT
5	50	2.00	0.7	0.6	4.00	23	0.5	2.80	120	0.3
5	100	1.00	0.8	0.6	2.00	27	0.5	1.40	124	0.3
5	200	0.50	1.1	0.8	1.00	40	0.6	0.70	123	0.3
5	500	0.20	2.0	1.4	0.40	76	1.2	0.28	140	0.5
5	1000	0.10	3.7	2.4	0.20	114	2.3			
5	2500	0.04	6.6	4.2	0.08	199	4.2			
10	50	1.00	0.8	0.6	2.00	24	0.4	1.40	117	0.3
10	100	0.50	1.1	0.7	1.00	28	0.6	0.70	102	0.3
10	200	0.25	1.7	1.0	0.50	44	1.0	0.35	121	0.5
10	500	0.10	3.6	2.3	0.20	111	2.3	0.14	155	0.7
10	1000	0.05	6.1	3.8	0.10	195	3.8			
10	2500	0.02	12.6	7.8	0.04	400	8.0			
25	50	0.40	1.1	0.7	0.80	38	0.6	0.56	118	0.3
25	100	0.20	1.7	1.1	0.40	70	1.2	0.28	137	0.5
25	200	0.10	3.7	2.2	0.20	118	2.3	0.14	155	0.7
25	500	0.04	6.6	4.4	0.08	200	4.2	0.06	242	1.6
25	1000	0.02	12.5	8.3	0.04	389	7.8			
25	2500	0.01	30.6	19.6	0.02	966	19.7			

50 Monte Carlo simulations

to the true value and only fails to contain the true value by a small amount of 0.01 or less. Therefore, the estimated intervals of both the Horowitz–Manski and Manski–Tamer estimators appear to be valid.

The pattern is similar with 10 missing observations, as summarized in Table 6. The estimated intervals of the Horowitz–Manski and Manski–Tamer approaches tend to contain the true parameter value for the parameters *k*, *b* and *d*, and the intervals—while large with small samples such as 50 groups—tend to fall quickly as the sample sizes grow. The Manski–Tamer approach provides very similar bounds, except for low sample sizes such as 50 groups with a group size of 5 members. All the estimated intervals are bigger than in the case of the 5 missing observations, as expected. For the *J* parameter of endogenous social interactions, the intervals can be quite big in small sample sizes with just 50 and 100 groups, even for groups with 25 members. It is also found that the estimated intervals do not contain the true *J* parameter for the cases of samples with 1000 and 2500 groups, although the width of the intervals falls with the sample size. In general, all the estimated intervals are larger with 10 missing values (in Table 6) relative to just 5 missing values (Table 5) as expected, but with bigger differences for the small samples such as 50 and 100 groups.

6 Conclusions and possible extensions

This paper examines partial inference of the peer effects models in the presence of missing outcome data, with a special focus on the binary choice case. Most peer effects models use the average outcome of each group as an explanatory variable; therefore, missing outcome data imply that we face both a problem of missing outcome values and an undetermined regressor. Having information on the bounds of the outcome variable can, however, help us get partial identification bounds for the parameters (Manski & Tamer, 2002, Horowitz & Manski, 2006). I use this information to obtain identification of a family of parametric binary choice models with peer effects (Brock & Durlauf, 2002; 2007, Blume et al., 2010), although a similar approach can be suggested for the linear peer effects model for the case in which identification can be obtained through non-closed peer groups. Other extensions of these results can easily be made by including a more general multinomial setting or semi-parametric discrete choice peer effect models (Blume et al., 2010).

Table 5 Minimum and maximum bounds around the true coefficients of the Horowitz–Manski and Manski–Tamer estimators of the logit endogenous social interactions model with non-closed groups and individuals excluded from their own peer group

Group size	No. of groups	HM								MT							
		$\bar{\theta}_{\min} - \theta$				$\bar{\theta}_{\max} - \theta$				$\bar{\theta}_{\min} - \theta$				$\bar{\theta}_{\max} - \theta$			
		k	*b*	*d*	*J*	*k*	*b*	*d*	*J*	*k*	*b*	*d*	*J*	*k*	*b*	*d*	*J*
5	50	−0.07	−0.16	−0.10	−0.98	0.32	0.01	0.06	0.96	−0.07	−0.16	−0.10	−1.79	0.43	0.07	0.08	0.96
5	100	0.02	−0.06	−0.07	−0.91	0.22	0.04	0.01	0.01	0.02	−0.09	−0.07	−0.91	0.22	0.04	0.02	0.01
5	200	0.04	−0.06	−0.02	−0.41	0.15	−0.01	0.02	0.08	−0.01	−0.06	−0.02	−0.41	0.15	0.03	0.03	0.15
5	500	0.00	0.00	−0.01	−0.06	0.04	0.02	0.00	0.12	0.00	−0.01	−0.01	−0.06	0.04	0.02	0.00	0.14
5	1000	0.01	0.00	0.00	0.01	0.03	0.01	0.00	0.11	0.01	0.00	−0.01	−0.03	0.04	0.01	0.00	0.11
5	2500	0.02	0.00	−0.01	−0.01	0.03	0.00	−0.01	0.03	0.02	0.00	−0.01	−0.01	0.03	0.00	0.00	0.07
10	50	−0.14	−0.03	−0.01	−0.80	0.12	0.07	0.08	0.54	−0.14	−0.05	−0.01	−0.91	0.22	0.07	0.08	0.54
10	100	−0.05	−0.01	−0.03	−0.18	0.08	0.05	0.01	0.47	−0.05	−0.01	−0.03	−0.18	0.08	0.05	0.01	0.47
10	200	−0.02	0.00	−0.01	−0.16	0.04	0.02	0.01	0.15	−0.02	−0.02	−0.01	−0.16	0.05	0.02	0.01	0.15
10	500	−0.04	0.00	−0.01	0.17	−0.01	0.01	0.00	0.29	−0.04	−0.01	−0.02	0.12	0.01	0.01	0.00	0.29
10	1000	0.00	−0.01	−0.01	0.07	0.01	0.00	0.00	0.13	−0.01	−0.01	−0.01	0.07	0.01	0.01	0.00	0.17
10	2500	−0.01	0.00	−0.01	0.14	0.00	0.00	0.00	0.16	−0.01	−0.01	−0.01	0.14	0.00	0.00	0.00	0.17
25	50	−0.04	−0.04	−0.04	−0.25	0.07	0.00	0.00	0.36	−0.04	−0.04	−0.04	−0.72	0.15	0.02	0.02	0.36
25	100	0.01	−0.01	0.00	−0.29	0.06	0.01	0.01	−0.01	0.01	−0.01	−0.01	−0.29	0.06	0.01	0.01	−0.01
25	200	0.00	0.00	−0.02	−0.07	0.03	0.01	−0.01	0.07	0.00	0.00	−0.02	−0.07	0.03	0.01	0.01	0.07
25	500	−0.01	−0.01	0.00	0.06	0.00	0.00	0.00	0.12	−0.03	−0.01	0.00	0.06	0.00	0.00	0.00	0.23
25	1000	−0.03	0.00	0.00	0.20	−0.03	0.00	0.00	0.23	−0.03	0.00	−0.02	0.19	−0.02	0.00	0.00	0.23
25	2500	−0.01	0.00	−0.01	0.10	−0.01	0.00	0.00	0.11	−0.01	0.00	−0.01	0.10	−0.01	0.00	0.00	0.11

50 Monte Carlo simulations, 5 missing values in each simulation

For the case of bounded variables, sharp bounds can be obtained for all group variables and outcomes by plugging in all possible combination of values of the missing variables (Horowitz & Manski, 2006). This method, however, is computationally difficult to implement, since the number of potential combinations increases exponentially with the number of groups and therefore quickly becomes a heavy computational exercise even for datasets of moderate size. An attractive alternative, however, can be developed by noticing this model has an interval (I), monotonicity (M) and mean independence (MI) properties, which can be summarized jointly as the IMMI assumption. Using these properties, a modified minimum distance (MMD) estimator is presented to obtain non-sharp bounds for the coefficients. While this approach is here suggested as a solution to the binary peer effects case, the same estimator can be easily applied to any parametric model with missing outcomes and interval regressors. In a set of Monte Carlo exercises, I show that the non-sharp bounds obtained through an interval estimator similar to Manski and Tamer (2002) provide results quite similar to the sharp bounds of the Horowitz and Manski (2006) approach, but at a much smaller cost in terms of computational time. The computational time of the Horowitz and Manski (2006) approach increases exponentially with the number of missing observations and can quickly become overwhelming with just 15 missing outcomes, but the non-sharp bounds proposed as an alternative with the IMMI assumption do not increase their computational time with additional missing outcomes and provide a good approximation for the sharp intervals (at least for the calibrated Monte Carlo exercises considered in this article). The Monte Carlo exercises also show that for the binary discrete choice model of peer effects there is not a significantly higher estimation accuracy for the case of non-closed groups relative to the closed groups case.

The bounds of the interval estimators of peer effects in the specified exercises are still large. This is a case for future econometricians and applied economists to combine further realistic assumptions in order to obtain tighter bounds (Manski, 2003).

Appendix 1: Proofs

Proof of Propositions 1 and 2

Let ν be any interval-valued variable with $\nu \in [\nu_0, \nu_1]$ (Assumption I). Let $E[y \mid x, \nu]$ be weakly increasing in ν (monotonicity—Assumption M). The law of iterated expectations and assumption mean independence (MI: $E[y \mid x, \nu, \nu_0, \nu_1] = E[y \mid x, \nu]$) yield

Table 6 Minimum and maximum bounds around the true coefficients of the Horowitz–Manski and Manski–Tamer estimators of the logit endogenous social interactions model with non-closed groups and individuals excluded from their own peer group

Group size	No. of groups	HM								MT							
		$\bar{\theta}_{\min} - \theta$				$\bar{\theta}_{\max} - \theta$				$\bar{\theta}_{\min} - \theta$				$\bar{\theta}_{\max} - \theta$			
		k	***b***	***d***	***J***	***k***	***b***	***d***	***J***	***k***	***b***	***d***	***J***	***k***	***b***	***d***	***J***
5	50	−0.12	−0.29	−0.16	−2.51	0.63	0.09	0.17	1.03	−0.22	−0.29	−0.16	−2.51	0.63	0.09	0.17	1.20
5	100	−0.12	−0.13	−0.11	−0.94	0.29	0.05	0.03	0.93	−0.12	−0.13	−0.11	−0.94	0.29	0.05	0.03	0.93
5	200	−0.03	−0.05	−0.05	−0.60	0.16	0.04	0.03	0.34	−0.03	−0.05	−0.05	−0.60	0.16	0.04	0.03	0.34
5	500	0.00	−0.02	−0.03	−0.25	0.08	0.01	0.00	0.12	0.00	−0.02	−0.03	−0.25	0.08	0.01	0.00	0.21
5	1000	−0.01	−0.01	−0.01	0.00	0.03	0.01	0.01	0.19	−0.01	−0.01	−0.01	0.00	0.03	0.01	0.01	0.19
5	2500	0.02	0.00	−0.01	0.00	0.03	0.01	0.00	0.07	0.02	−0.01	−0.01	0.00	0.03	0.01	0.00	0.08
10	50	−0.03	−0.14	−0.08	−1.83	0.44	0.04	0.09	0.55	−0.03	−0.14	−0.08	−1.83	0.44	0.05	0.09	0.55
10	100	−0.05	−0.09	−0.05	−0.58	0.18	0.01	0.03	0.58	−0.05	−0.09	−0.05	−0.58	0.19	0.01	0.03	0.58
10	200	−0.05	−0.03	−0.04	−0.13	0.06	0.01	0.00	0.46	−0.05	−0.04	−0.04	−0.13	0.06	0.01	0.01	0.46
10	500	0.00	−0.02	−0.02	−0.07	0.05	0.00	0.00	0.17	−0.01	−0.02	−0.02	−0.07	0.05	0.00	0.00	0.19
10	1000	−0.02	−0.01	−0.01	0.10	0.00	0.00	−0.01	0.22	−0.02	−0.01	−0.01	0.10	0.00	0.00	−0.01	0.22
10	2500	−0.02	−0.01	−0.01	0.18	−0.01	0.00	−0.01	0.23	−0.03	−0.01	−0.01	0.18	−0.01	0.00	−0.01	0.23
25	50	−0.05	−0.08	−0.04	−0.78	0.17	0.00	0.03	0.40	−0.05	−0.08	−0.04	−0.78	0.17	0.01	0.03	0.40
25	100	−0.03	−0.03	−0.03	−0.25	0.07	0.01	0.00	0.30	−0.03	−0.03	−0.03	−0.25	0.07	0.01	0.01	0.30
25	200	−0.04	−0.02	−0.01	−0.05	0.02	0.00	0.01	0.24	−0.04	−0.02	−0.01	−0.09	0.03	0.01	0.01	0.24
25	500	−0.02	0.00	−0.01	0.04	0.00	0.00	0.00	0.15	−0.02	−0.02	−0.01	−0.05	0.02	0.00	0.00	0.15
25	1000	−0.03	0.00	−0.01	0.15	−0.02	0.00	−0.01	0.21	−0.03	0.00	−0.01	0.15	−0.02	0.00	0.00	0.21
25	2500	−0.02	0.00	−0.01	0.16	−0.02	0.00	0.00	0.18	−0.02	0.00	−0.01	0.13	−0.02	0.00	0.00	0.18

50 Monte Carlo simulations, 10 missing values in each simulation

$$E[y \mid x, \nu_0, \nu_1] = \int E[y \mid x, \nu, \nu_0, \nu_1] \partial P(\nu \mid x, \nu_0, \nu_1] = \int E[y \mid x, \nu] \partial P(\nu \mid x, \nu_0, \nu_1] \tag{A1}$$

where the first equality is given by the law of iterated expectations and the second one by Assumption MI.

Assumptions I and M imply that for all constants $V_0 \leq V_1$,

$$E_L[y \mid x, \nu = V_0] \leq E[y \mid x, \nu = V_0] \tag{A.3.1}$$

$$E[y \mid x, \nu = V_1] \leq E_U[y \mid x, \nu = V_1] \tag{A.3.2}$$

$$E_L[y \mid x, \nu_0 = V_0, \nu_1 = V_1] \leq E[y \mid x, \nu_0 = V_0, \nu_1 = V_1] \leq E_L[y \mid x, \nu_0 = V_0, \nu_1 = V_1] \tag{A.3.3}$$

where

$$E_L[y \mid x, \nu = V_0] = E[y \mid x, \nu = V_0, z = 1] P(z = 1 \mid x, \nu = V_0) + y_L P(z = 0 \mid x, \nu = V_0)$$
$$E_U[y \mid x, \nu = V_1] = E[y \mid x, \nu = V_1, z = 1] P(z = 1 \mid x, \nu = V_1) + y_U P(z = 0 \mid x, \nu = V_1)$$
$$E_L[y \mid x, \nu_0 = V_0, \nu_1 = V_1] = E[y \mid x, \nu_0 = V_0, \nu_1 = V_1, z = 1] P(z = 1 \mid x, \nu_0 = V_0, \nu_1 = V_1) + y_L P(z = 0 \mid x, \nu_0 = V_0, \nu_1 = V_1)$$
$$E_U[y \mid x, \nu_0 = V_0, \nu_1 = V_1] = E[y \mid x, \nu_0 = V_0, \nu_1 = V_1, z = 1] P(z = 1 \mid x, \nu_0 = V_0, \nu_1 = V_1) + y_U P(z = 0 \mid x, \nu_0 = V_0, \nu_1 = V_1)$$

$$E[y \mid x, \nu = V_0] \leq \int E[y \mid x, \nu] \partial P(\nu \mid x, \nu_0 = V_0, \nu_1 = V_1] \leq E[y \mid x, \nu = V_1] \tag{A2}$$

and assumption that $y \in [y_L, y_U]$ and the law of total probability give when there are missing outcome data,

Hence,

$$E_L[y \mid x, \nu = V_0] \leq E[y \mid x, \nu_0 = V_0, \nu_1 = V_1] \leq E_U[y \mid x, \nu = V_1]. \tag{A4}$$

To prove the lower bound on $E[y \mid x, \nu = V]$, take any $V_1 \leq V$. It follows from A3 and from Assumption M that

$$E_L[y \mid x, v_0 = V_0, v_1 = V_1] \leq E[y \mid x, v_0 = V_0, v_1 = V_1] \leq E[y \mid x, v = V] \Rightarrow E_L[y \mid x, v_0 = V_0, v_1 = V_1] \leq E[y \mid x, v = V].$$

Hence, the lower bound holds. To prove sharpness, view the bound as a function of V. This function is weakly increasing in V, so Assumption M holds. The proof of the sharp upper bound uses analogous reasoning. Therefore, we have proved that under Assumption IMMI, we have:

$$\sup_{v_1 \leq V} E_L[y \mid x, v_0 = V_0, v_1 = V_1] \leq E[y \mid x, v = V] \leq \inf_{v_o \geq V} E_U[y \mid x, v_0 = V_0, v_1 = V_1].$$

In the absence of other information, these bounds are sharp. Propositions 1 and 2 are just special cases of this result.

Proof for Proposition 3

Assumption IMMI gives us

$$E[y \mid x, v = V_0] \leq E[y \mid x, v_0 = V_0, v_1 = V_1] \leq E[y \mid x, v = V_1]. \tag{A5}$$

For a parametric model, this inequality becomes

$$f(x, v_0, \gamma) \leq E[y \mid x, v_0, v_1] \leq f(x, v_1, \gamma). \tag{A6}$$

For the case of missing outcome data, $E[y \mid x, v_0, v_1]$ is not perfectly observed, but A.3.3 gives us $E_L[y \mid x, v_0, v_1] \leq E[y \mid x, v_0, v_1] \leq E_U[y \mid x, v_0, v_1]$. A6 and A3 give us $f(x, v_0, \gamma) \leq E_U[y \mid x, v_0, v_1]$ and $f(x, v_1, \gamma) \geq E_L[y \mid x, v_0, v_1]$. Therefore, for a parametric model, inequalities A6 and A3 become:

$$f(x, v_0, \gamma) \leq E_U[y \mid x, v_0, v_1] \bigcup f(x, v_1, \gamma) \geq E_L[y \mid x, v_0, v_1]. \tag{A7}$$

It follows that c is equivalent to γ if and only if $f(x, v_0, c) \leq E_U[y \mid x, v_0, v_1] \bigcup f(x, v_1, c) \geq E_L[y \mid x, v_0, v_1]$, a.e. (x, v_0, v_1) . □

Proof for Lemma 1

This corollary allows us to characterize the identification region for the case of the monotone index form $f(x, v, \gamma) = F(x\beta + \delta v)$ in the case of missing outcome data. The identification region of γ in our Proposition 3 is given by $C^* \equiv \{c \in C : P(V(c)) = 0\}$, where $V(c) \equiv [\bigcup (x, v_0, v_1) : f(x, v_0, c) \leq E_U[y \mid x, v_0, v_1] \; f(x, v_1, c) \geq E_L[y \mid x, v_0, v_1]]$, as proved previously.

Let f have the monotone-index form. Then:

(a) C^* is non-empty and convex.

(b) Assume that there exists no proper linear subspace of R^{k+1} having probability one under $P(x, v)$. Assume that $P(v_0 = v_1) > 0$ and $P(z = 1 \mid v_0 = v_1) = 1$. Then, $C^* = \gamma$.

Proof (a) The set C^* is non-empty because $\gamma \in C^*$. To prove convexity, observe that the condition $P(V(c)) = 0$, identifying $c = (b, d)$ as a member of C^*, holds if and only if

$$P(F(x, v_0, c) \leq E_U[y \mid x, v_0, v_1] \bigcup F(x, v_1, c) \geq E_L[y \mid x, v_0, v_1]) = P(xb + dv_0 \leq s_U(x, v_0, v_1) \bigcup xb + dv_1 \geq s_L(x, v_0, v_1)) = 1$$

where $s_U(x, v_0, v_1) \equiv F^{-1}(E_U[y \mid x, v_0, v_1])$ and $s_L(x, v_0, v_1) \equiv F^{-1}(E_L[y \mid x, v_0, v_1])$. Let c' and c'' be distinct elements of C^*. Then,

$$P(xb' + d'v_0 \leq s_U(x, v_0, v_1) \bigcup xb' + d'v_1 \geq s_L(x, v_0, v_1)) = P(xb'' + d''v_0 \leq s_U(x, v_0, v_1) \bigcup xb'' + d''v_1 \geq s_L(x, v_0, v_1)) = 1.$$

Now consider $c_\alpha \equiv \alpha c' + (1 - \alpha)c''$, where $\alpha \in (0, 1)$. It follows from the above that

$$P(xb_\alpha + d_\alpha v_0 \leq s_U(x, v_0, v_1) \bigcup xb_\alpha + d_\alpha v_1 \geq s_L(x, v_0, v_1)) = 1.$$

Hence, $c_\alpha \in C^*$.

(b) Consider the subpopulation with $(v_0 = v_1)$. By assumption $P(v_0 = v_1) > 0$ and $P(z = 1 \mid v_0 = v_1) = 1$. Hence, $c \in C^*$ must satisfy the inequality $F(xb + dv) = E_U[y \mid x, v_0, v_1] = E_L[y \mid x, v_0, v_1] = E[y \mid x, v_0, v_1]$ or equivalently $xb + dv = F^{-1}(E[y \mid x, v_0, v_1]) = s(x, v_0, v_1)$, a.e. $(v_0 = v_1)$. The support condition on $P(x, v)$ implies that (β, δ) is the only parameter value that satisfies the equality almost everywhere $(v_0 = v_1)$. Hence, γ is identified. □

Result (b) is equivalent to saying that we are able to point-identify the parameters of the social interactions models if there is at least one group with no missing data. This is obviously a very strong to use in practice. Even if there is one or more groups with no missing data, we would need the sample size represented by these groups with no missing data to increase to infinity in order to avoid sampling imprecision in the estimation of the parameters.

Proof for Proposition 4

Let the estimator for the identification region be given by

$$H_N(\gamma) = \arg\min_{c\in C} \frac{1}{N}\sum_{i=1}^{N} g_{N1}(c, x_i, v_{0i}, v_{1i})\ w_{N1}(c, x_i, v_{0i}, v_{1i}) + g_{N0}(c, x_i, v_{0i}, v_{1i})\ \ w_{N0}(c, x_i, v_{0i}, v_{1i})$$

where

$$g_{N0}(c, x_i, v_{0i}, v_{1i}) = 1[f(x_i, v_{0i}, c) > E_U^N[y \mid x_i, v_{0i}, v_{1i}]],$$
$$g_{N1}(c, x_i, v_{0i}, v_{1i}) = 1[f(x_i, v_{1i}, c) < E_L^N[y \mid x_i, v_{0i}, v_{1i}]],$$
$$w_{N0}(c, x_i, v_{0i}, v_{1i}) = [f(x_i, v_{0i}, c) - E_U^N[y \mid x_i, v_{0i}, v_{1i}]]^2 \text{and}$$
$$w_{N1}(c, x_i, v_{0i}, v_{1i}) = [f(x_i, v_{1i}, c) - E_L^N[y \mid x_i, v_{0i}, v_{1i}]]^2,$$

with $E_L^N[y \mid x_i, v_{0i}, v_{1i}]$ and $E_U^N[y \mid x_i, v_{0i}, v_{1i}]$ being consistent estimates of $E_L[y \mid x_i, v_{0i}, v_{1i}]$ and $E_U[y \mid x_i, v_{0i}, v_{1i}]$.

Proof Manski and Tamer (2002) provide a proof that $H_N(\gamma)$ is a consistent estimator for the identification region $H(\gamma)$, which remains valid in this case with $E_U^N[y \mid x_i, v_{0i}, v_{1i}]$ and $E_L^N[y \mid x_i, v_{0i}, v_{1i}]$ in the place of $\eta_N(x_i, v_{0i}, v_{1i}) = E^N[y \mid x_i, v_{0i}, v_{1i}]$, and therefore the proof is omitted here.

Appendix 2: Monte Carlo simulations for the linear model

Interval estimators for the linear case

For the linear social interactions case, I will assume an identified model in which the peer group is non-closed:

$$y_i = k + bX_i + dY_{g(i)} + J(\bar{y}_{i,g} + w_i) + \epsilon_i, \qquad \text{(B.1)}$$

with all the regressors $(X_i, Y_{g(i)}, \bar{y}_{i,g} + w_i)$ and the unobserved error term ϵ_i being bounded. $\bar{y}_{i,g}$ represents the average outcomes among the peer group in the sample, while w_i represents the average outcomes of other peers of individual i but which are not peers of the other members of group $g(i)$. I assume that both $\bar{y}_{i,g}$ and w_i are observed (for instance, the individuals could self-report the average outcomes of their other peers which are not common peers in the group $g(i)$). Since all the terms $W_i \equiv (\bar{y}_{i,g} + w_i, X_i, Y_{g(i)}, \epsilon_i)$ are bounded, that is $W_i \in [v_L, v_U]$ with v_L, v_U being multivariate vectors of size $2 + K + Q$, the outcomes are bounded in an interval as well: $y \in [y_L, y_U]$.

Again assume y is observed when $z = 1$ and not observed when $z = 0$. For simplicity, I assume that X_i, $Y_{g(i)}$ and the outside peer effect of the individual w_i are always observed, but the missing information on some outcomes y_i implies that $\bar{y}_{i,g}$ is not point-identified. Again, I denote $V_x^g = (E(y \mid g), w, Y_g, x)$, $W_x^g = (E_L(y \mid g), E_U(y \mid g), w, Y_g, x)$, $W_0^g = (E_L(y \mid g), w, Y_g, x)$, and $W_1^g = (E_U(y \mid g), w, Y_g, x)$. This is similar to the previous definition, which used $p_{g(i)}, p_{g(i)}^L, p_{g(i)}^U$ instead of $E(y \mid g), E_L(y \mid g), E_U(y \mid g)$. I also assume the standard location assumption, $E[\epsilon_i \mid (\bar{y}_{i,g} + w_i), Y_g, x] = 0$.

This linear social interactions model complies with the IMMI assumptions, just like the previously exposed discrete choice model. In particular, the linear social interactions model satisfies: i) the interval assumption (I), because $y \in [y_L, y_U]$ and $\bar{y}_{i,g} \in [y_L, y_U]$; ii) the weak monotonicity assumption (M), since $E[y \mid W_{x'}^{g'}]$ is weakly increasing in $\bar{y}_{i,g}$ due to J being a constant; iii) the mean independence assumption (MI), since $E[y \mid E(y \mid g), W_x^g] = E[y \mid V_x^g] = k + bX_i + dY_{g(i)} + J(\bar{y}_{i,g} + w_i)$.

Assumption I) and the law of total probability give us Proposition 2.B: $E_L[y \mid W_x^g] \le E[y \mid W_x^g] \le E_U[y \mid W_x^g]$, where $E_L[y \mid W_x^g] = E[y \mid W_x^g, z = 1] P(z = 1 \mid W_x^g) + y_L P(z = 0 \mid W_x^g)$ and $E_U[y \mid W_x^g] = E[y \mid W_x^g, z = 1] P(z = 1 \mid W_x^g) + y_U P(z = 0 \mid W_x^g)$. This is similar to Proposition 2, which applied $y_L = 0$ and $y_U = 1$.

Then, by assumptions IMMI we get Proposition 4.B: Let $G(W_i) = k + bX_i + dY_{g(i)} + J(\bar{y}_{i,g} + w_i)$. A suggested estimator for the identification region $H(\theta)$ would be

$$H_N(\theta) = \arg\min_{c\in C} \frac{1}{N}\sum_{i=1}^{N} 1[G(W_{1,i}^g) < E_L^N[y \mid W_{x,i}^g]]\ [G(W_{1,i}^g) - E_L^N[y \mid W_{x,i}^g]]^2 + 1[G(W_{0,i}^g) > E_U^N[y \mid W_{x,i}^g]]\ [G(W_{0,i}^g) - E_U^N[y \mid W_{x,i}^g]]^2$$

where $E_L^N[y \mid W_{x,i}^g]$ and $E_U^N[y \mid W_{x,i}^g]$ are consistent estimators of $E_L[y \mid W_{x,i}^g]$ and $E_U[y \mid W_{x,i}^g]$, respectively.

Monte Carlo exercises

The linear peer effects model is simulated as follows. For each simulation s, the model is given by $y_i = k + bX_i + dY_{g(i)} + J(\bar{y}_{i,g} + w_i) + \epsilon_i$. Again, the simulations consider that the coefficients $\{k, b, d, J\}$ are constants for all simulations and that all the groups are the same, that is, $n_{g(i)} = n_g$. Each simulated observation is obtained with the specified set of coefficients: $k = 1.5$, $b = 0.5$, $d = 0.3$, $J = 0.2$. The variables $X_i(s)$, $Y_{g(i)}(s)$, $w_i(s)$ and $\varepsilon_i(s)$ are simulated as independent

Table 7 Bias, standard deviation and mean absolute deviation of the estimates for the linear (OLS) model with non-closed groups

Group size	No. of groups	Bias: $\bar{\theta} - \theta(\bar{\theta} = \frac{\sum_S \hat{\theta}_s}{S})$				STD: $\sqrt{\frac{\sum_S (\hat{\theta}_s - \bar{\theta})^2}{S-1}}$				MAD: $\frac{\sum_S \lvert\hat{\theta}_s - \theta\rvert}{S}$				Average time (secs)
		k	*b*	*d*	*J*	*k*	*b*	*d*	*J*	*k*	*b*	*d*	*J*	
OLS (non-closed groups)														
5	50	−0.01	0.00	0.00	0.00	0.12	0.01	0.08	0.04	0.09	0.01	0.06	0.03	0.02
5	100	−0.01	0.00	−0.01	0.00	0.08	0.01	0.05	0.03	0.07	0.01	0.04	0.02	0.01
5	200	−0.01	0.00	−0.01	0.01	0.06	0.01	0.03	0.02	0.05	0.01	0.02	0.02	0.01
5	500	−0.01	0.00	0.00	0.00	0.04	0.00	0.02	0.01	0.03	0.00	0.02	0.01	0.02
10	50	0.00	0.00	0.00	0.00	0.09	0.01	0.05	0.03	0.07	0.01	0.04	0.02	0.01
10	100	0.00	0.00	0.00	0.00	0.07	0.01	0.03	0.02	0.06	0.01	0.02	0.02	0.01
10	200	−0.01	0.00	0.00	0.00	0.04	0.01	0.02	0.01	0.03	0.00	0.02	0.01	0.01
10	500	0.00	0.00	0.00	0.00	0.03	0.00	0.01	0.01	0.02	0.00	0.01	0.01	0.02
25	50	0.01	0.00	0.00	0.00	0.06	0.01	0.03	0.02	0.05	0.01	0.03	0.02	0.01
25	100	−0.01	0.00	0.00	0.00	0.04	0.01	0.02	0.01	0.03	0.01	0.02	0.01	0.01
25	200	−0.01	0.00	0.00	0.00	0.03	0.00	0.02	0.01	0.02	0.00	0.01	0.01	0.02
25	500	−0.01	0.00	0.00	0.00	0.02	0.00	0.01	0.01	0.02	0.00	0.01	0.00	0.03

50 Monte Carlo simulations. Individuals excluded from their own peer group

Table 8 Time performance of the Horowitz–Manski (HM) and Manski–Tamer (MT) interval estimators for the linear peer effects models (with non-closed groups and individuals excluded from their own peer group). 50 Monte Carlo simulations

Group size	No. of groups	5 missing observations		
		Prob. of missing (in %)	Average time (secs)	
			OLS	
			HM	MT
5	50	2.00	6.8	0.3
5	100	1.00	6.9	0.4
5	200	0.50	6.9	0.4
5	500	0.20	8.2	0.6
10	50	1.00	6.7	0.4
10	100	0.50	6.5	0.4
10	200	0.25	7.0	0.6
10	500	0.10	9.1	0.8
25	50	0.40	6.4	0.4
25	100	0.20	6.9	0.5
25	200	0.10	8.2	0.8
25	500	0.04	12.6	1.7

pseudo-standard normal numbers, with respective supports between $[0, 4]$, $[0, 1]$, $[1.25, 4.125]$ and $[-0.5, 0.5]$. The reason why $w_i(s)$ is expressed between 1.25 and 4.125 is to match the same support as the value of $\sum_{j=1, j\in g(i)}^{n} 1(y_j(s) = 1)/n_g$. For the variable $p_{i,g}(s)$, I implement two alternatives:

(i) non-closed groups with individual i as part of his own peer group and with the outside peer group of the same size as the group g,
$$p_{i,g}(s) = \frac{n_g w_i(s) + \sum_{j=1, j\in g(i)}^{n} 1(y_j(s) = 1)}{n_g + n_g};$$

(ii) non-closed groups with individual i excluded from their own peer group and with the outside peer group of the same size as the group g,
$$p_{i,g}(s) = \frac{n_g w_i(s) + \sum_{j=1, j\in g(i), j\neq i}^{n} 1(y_j(s) = 1)}{n_g + n_g - 1}.$$

The reason why the OLS peer effects models do not consider closed groups is due to the well-known identification problem of including endogenous effects in linear models with closed groups (Manski 1993, Bramoullé et al., 2009), and therefore the peer effects $w_i(s)$ that are specific to each individual i are required for the identification. For the OLS model, I will consider the cases of 5 and 7 missing values, with the number of possible V outcomes being taken from a grid of 5 values: 1.25, 2.0, 2.7, 3.4 and 4.125. Specifying $V = 5$ in the linear case is an approximation, since in fact the outcome y is continuous and would require an infinite number of possible values for each outcome. Therefore, the linear case presents a lower bound for the computational demands of applying the Horowitz–Manski estimator.

Table 7 shows the performance of the linear endogenous peer effects model with non-closed groups (which is required for the identification). For simplicity, I only present the results with individuals excluded from their

Table 9 Minimum and maximum bounds around the true coefficients of the Horowitz–Manski and Manski–Tamer estimators of the OLS endogenous social interactions model with non-closed groups and individuals excluded from their own peer group. 50 Monte Carlo simulations, 5 missing values in each simulation

Group size	No. of groups	HM								MT							
		$\bar{\theta}_{\min} - \theta$				$\bar{\theta}_{\max} - \theta$				$\bar{\theta}_{\min} - \theta$				$\bar{\theta}_{\max} - \theta$			
		k	*b*	*d*	*J*	*k*	*b*	*d*	*J*	*k*	*b*	*d*	*J*	*k*	*b*	*d*	*J*
5 missing values																	
5	50	−0.13	−0.03	−0.09	−0.05	0.17	0.01	0.07	0.05	−0.13	−0.03	−0.09	−0.05	0.17	0.01	0.07	0.05
5	100	−0.09	−0.02	−0.06	−0.02	0.07	0.01	0.02	0.03	−0.09	−0.02	−0.06	−0.02	0.07	0.01	0.02	0.03
5	200	−0.05	−0.01	−0.03	−0.01	0.03	0.00	0.00	0.02	−0.05	−0.01	−0.03	−0.01	0.03	0.00	0.01	0.02
5	500	−0.02	−0.01	−0.01	0.00	0.01	0.00	0.01	0.01	−0.02	−0.01	−0.01	0.00	0.01	0.00	0.01	0.01
10	50	−0.05	−0.02	−0.04	−0.03	0.12	0.00	0.03	0.03	−0.05	−0.02	−0.04	−0.03	0.12	0.00	0.03	0.03
10	100	−0.03	0.00	−0.02	−0.02	0.06	0.01	0.01	0.01	−0.03	0.00	−0.02	−0.02	0.06	0.01	0.02	0.01
10	200	−0.02	0.00	−0.01	−0.01	0.01	0.00	0.01	0.01	−0.02	0.00	−0.01	−0.01	0.01	0.00	0.01	0.01
10	500	−0.01	0.00	0.00	−0.01	0.01	0.00	0.01	0.00	−0.01	0.00	0.00	−0.01	0.01	0.00	0.01	0.00
25	50	−0.06	0.00	−0.01	0.00	0.00	0.01	0.02	0.02	−0.06	−0.01	−0.01	−0.01	0.02	0.01	0.02	0.02
25	100	−0.01	−0.01	0.00	−0.01	0.03	0.00	0.02	0.00	−0.01	−0.01	−0.01	−0.01	0.03	0.00	0.02	0.00
25	200	−0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	−0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00
25	500	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7 missing values																	
5	50	−0.15	−0.04	−0.16	−0.09	0.30	0.02	0.08	0.06	−0.15	−0.04	−0.16	−0.09	0.30	0.02	0.08	0.06
5	100	−0.10	−0.02	−0.06	−0.04	0.13	0.01	0.05	0.03	−0.10	−0.02	−0.06	−0.04	0.13	0.01	0.05	0.03
5	200	−0.07	−0.01	−0.03	−0.01	0.03	0.00	0.02	0.02	−0.07	−0.01	−0.03	−0.01	0.03	0.00	0.02	0.02
5	500	−0.02	−0.01	−0.02	0.00	0.03	0.00	0.00	0.01	−0.02	−0.01	−0.02	0.00	0.03	0.00	0.01	0.01
10	50	−0.11	−0.02	−0.05	−0.03	0.11	0.01	0.06	0.04	−0.11	−0.02	−0.05	−0.03	0.11	0.01	0.06	0.04
10	100	−0.08	−0.01	0.00	−0.02	0.04	0.00	0.05	0.02	−0.08	−0.01	−0.01	−0.02	0.04	0.00	0.05	0.02
10	200	−0.02	−0.01	−0.01	−0.01	0.03	0.00	0.02	0.01	−0.02	−0.01	−0.01	−0.01	0.03	0.00	0.02	0.01
10	500	−0.01	0.00	−0.01	0.00	0.01	0.00	0.00	0.00	−0.01	0.00	−0.01	0.00	0.01	0.00	0.00	0.00
25	50	−0.03	−0.01	−0.04	−0.02	0.07	0.00	0.01	0.01	−0.03	−0.01	−0.04	−0.02	0.07	0.00	0.01	0.01
25	100	0.00	0.00	−0.02	−0.01	0.05	0.00	0.01	0.00	0.00	0.00	−0.02	−0.01	0.05	0.00	0.01	0.00
25	200	−0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00	−0.01	0.00	0.00	0.00	0.01	0.00	0.01	0.00
25	500	−0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	−0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

own peer effects, since otherwise there could be a significant bias in the estimation due to the correlation between $y_i(s)$ and the unobserved idiosyncratic error $\varepsilon_i(s)$. The results show that there is a very rapid convergence of the OLS estimates for all the coefficients even for sample sizes as small as 50 groups and a small group size of just 5 members. Therefore, in the case of non-closed groups, the convergence of the OLS estimator is much faster than for the logit model (model M3). Table 8 shows that for the linear model, the Horowitz–Manski type of estimator has an average performance time between 6.4 and 12.6 seconds for 5 missing observations, but this grows to an average time between 102 and 242 seconds with just 7 missing observations. However, the Manski–Tamer type of estimator keeps a similar time performance whether with 5 or 7 missing observations, with an average time between 0.3 and 1.7 seconds.

Finally, Table 9 shows the performance of the Horowitz–Manski and Manski–Tamer estimators for the linear peer effects model. In this case, both interval estimator approaches coincide perfectly, although perhaps this would not be the case with other calibrations or with a higher number of missing outcomes. It is possible that with a larger number of missing values, the interval estimates of the Manski–Tamer approach would be much worse than the sharp bounds of the Horowitz–Manski approach, although the Horowitz–Manski approach would certainly increase enormously its computational time due to the large number of possible missing datasets given by $MD = V^{N_{z=0}}$. In general, the interval estimates contain the true parameter value for all the coefficients, including the endogenous peer coefficient *J*. The intervals are somewhat wider when the missing observations increase from 5 to 7, as expected. But the estimated intervals of the linear model fall substantially and become

negligible for sample sizes of 200 groups or more. Therefore, the convergence of the intervals is much faster for the linear peer effects than in the discrete choice case. The Monte Carlo exercises show that, even in the case without any missing data, there are significant accuracy problems for estimating linear peer effects models that include the individuals as part of their own peer group, since this creates a problem of an endogenous regressor being correlated with the unobservable error term.

Abbreviations
I: Interval; IMMI: Interval, monotonicity, and mean independence; M: Monotonicity; MI: Mean independence; MLE: Maximum likelihood estimator.

Acknowledgements
The article benefited only from comments and suggestions of fellow academic researchers and seminar participants.

Author contributions
The author Carlos Madeira is the single author of this article and its contributions. The author read and approved the final manuscript.

Funding
This study was funded by Fundação Calouste Gulbenkian and Fundação para a Ciência e Tecnologia.

Availability of data and materials
The article does not use any source of data.

Declaration

Ethical approval
This article does not contain any studies with animals or humans performed by any of the authors.

Competing interests
The author Carlos Madeira declares that he has no conflict of interest or competing interests regarding this article.

Software and hardware used in the article
All the computational exercises in this work were performed with a Stata MP-6 version 15.1 software. The hardware used is an Intel Core i7-9750H CPU with 6 physical cores (12 virtual processors), 2.60 GHz of speed and 24.0 GB of available physical RAM.

Received: 30 October 2020 Accepted: 2 June 2022
Published online: 07 July 2022

References

Advani, A., & Malde, B. (2018). Methods to identify linear network models: a review. *Swiss Journal of Economics and Statistics, 154*, 12.
Ammermueller, A., & Pischke, J. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics, 27*(3), 15–348.
Bailey, M., D. Johnston, M. Koenen, T. Kuchler, D. Russel, & J. Stroebel (2021). Social Networks shape beliefs and behavior: Evidence from social distancing during the COVID-19 pandemic. NBER WP 28234.
Blume, L., Brock, W., Durlauf, S., & Ioannides, Y. (2010). Identification of social interactions. *Handbook of Social Economics, 18*, 853–964.
Bramoullé, Y., Djebbari, H., & Fortin, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics, 150*, 41–55.
Brock, W., & Durlauf, S. (2002). A multinomial-choice model of neighborhood effects. *American Economic Review, 92*(2), 298–303.
Brock, W., & Durlauf, S. (2007). Identification of binary choice models with social interactions. *Journal of Econometrics, 140*(1), 52–75.
Bruhin, A., Goette, L., Haenni, S., & Jiang, L. (2020). Spillovers of prosocial motivation: Evidence from an intervention study on blood donors. *Journal of Health Economics, 70*, 102244.
Chernozhukov, V., Hong, H., & Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica, 75*(5), 1243–1284.
Cipollone, P., & Rosolia, A. (2007). Social interactions in high school: Lessons from an earthquake. *American Economic Review, 97*(3), 948–965.
Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives, 15*(4), 57–67.
Hausman, J., Abrevaya, J., & Scott-Morton, F. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics, 87*(2), 239–269.
Heckman, J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In: C. F. Manski and D. McFadden (Eds.) Structural analysis of discrete panel data with econometric applications (pp. 179–195).
Honoré, B., & Tamer, E. (2006). Bounds on parameters in panel dynamic discrete choice models. *Econometrica, 74*(3), 611–629.
Horowitz, J., & Manski, C. (2006). Identification and estimation of statistical functionals using incomplete data. *Journal of Econometrics, 132*(2), 445–459.
Imbens, G., & Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica, 72*(6), 1845–1857.
Kooreman, P., & Soetevent, A. (2007). A discrete-choice model with social interactions: with an application to high school teen behavior. *Journal of Applied Econometrics, 22*, 599–624.
Krauth, B. (2006). Simulation-based estimation of peer effects. *Journal of Econometrics, 133*(1), 243–271.
Lalive, R., & Cattaneo, A. (2009). Social interactions and schooling decisions. *Review of Economics and Statistics, 91*(3), 457–477.
Madeira, C. (2018). Testing the rationality of expectations of qualitative outcomes. *Journal of Applied Econometrics, 33*(6), 837–852.
Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies, 60*(3), 531–542.
Manski, C. (2000). Economic analysis of Social Interactions. *Journal of Economic Perspectives, 14*(3), 115–136.
Manski, C. (2003). Partial Identification of Probability Distributions. Springer Series in Statistics.
Manski, C., & Tamer, E. (2002). Inference on regressions with interval data on a regressor or outcome. *Econometrica, 70*(2), 519–546.
Roth, A. (2020). How the provision of childcare affects attitudes towards maternal employment. *Swiss Journal of Economics and Statistics, 156*, 17.
Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics, 116*(2), 681–704.
Slotwinski, M., Stutzer, A., & Uhlig, R. (2019). Are asylum seekers more likely to work with more inclusive labor market access regulations? *Swiss Journal of Economics and Statistics, 155*, 17.
Sojourner, A. (2013). Identification of peer effects with missing peer data: Evidence from project STAR. *Economic Journal, 123*(569), 574–605.
Wooldridge, J. (2010). *Econometric analysis of cross section and panel data*. MIT Press.

Publisher's Note